Data Ladder
Get the most out of your data

Whitepaper
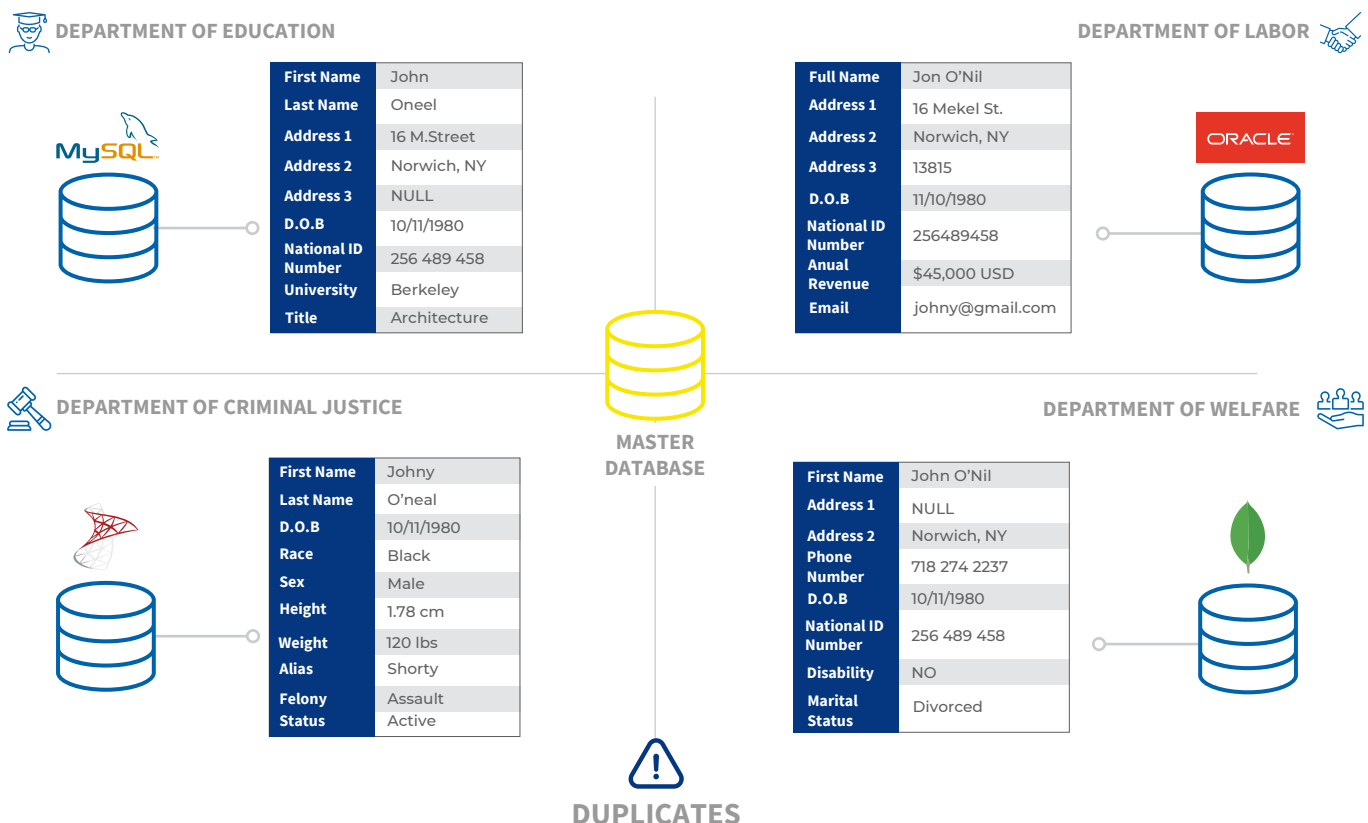
Identity Resolution for Government and Public Sector Institutions

Match records across disparate systems, datasets, and jurisdictions and enhance operational efficiency with identity resolution solution for government and public sector institutions.

Duplicate and inaccurate data records are a universal problem, however, when it comes to government databases, they can be the cause of damaging circumstances. There are examples everywhere.

Take, for example, **one state that carried out a sample** using an old program to analyze the number of students in one year who attended post-secondary education in a specific city. Using the outdated system, the state found that 22% out of 5,344 students in the city had gone on to higher education. When the state used Data Ladder's identity resolution software, the number went up to 41%. This example was just a small experiment conducted by the state to test the accuracy of two systems – imagine the consequences of using this flawed insight to obtain a grant or publish a study.

At Data Ladder, we have worked with the U.S Department of Education, Department of Transportation, Department of Treasury to ensure that government and public organizations are able to get a single view of truth through the implementation of various data cleansing and matching techniques.

DEPARTMENT OF EDUCATION

First Name	John
Last Name	Oneel
Address 1	16 M.Street
Address 2	Norwich, NY
Address 3	NULL
D.O.B	10/11/1980
National ID Number	256 489 458
University	Berkeley
Title	Architecture

DEPARTMENT OF LABOR

Full Name	Jon O'Nil
Address 1	16 Mekel St.
Address 2	Norwich, NY
Address 3	13815
D.O.B	11/10/1980
National ID Number	256489458
Anual Revenue	$45,000 USD
Email	johny@gmail.com

ORACLE

MASTER DATABASE

DEPARTMENT OF CRIMINAL JUSTICE

First Name	Johny
Last Name	O'neal
D.O.B	10/11/1980
Race	Black
Sex	Male
Height	1.78 cm
Weight	120 lbs
Alias	Shorty
Felony	Assault
Status	Active

DEPARTMENT OF WELFARE

First Name	John O'Nil
Address 1	NULL
Address 2	Norwich, NY
Phone Number	718 274 2237
D.O.B	10/11/1980
National ID Number	256 489 458
Disability	NO
Marital Status	Divorced

DUPLICATES

Identity Resolution as the Stepping Stone
to Data Quality Management

Data is the lifeline of any government organization. Here's another example of how poor data quality can lead to inaccurate information resulting in actual damages.

In California, **account records at the State Controller's Office** had instances where sick leaves and vacation credits were inaccurately recorded. Instead of 8 hours, there were fields where it was 80 and even 800. The system, being outdated, didn't have the controls to prevent this error. The audit found 200,000 inexplicable hours of leave due to data errors. Those leave hours had a value of $6 million.

Data quality, therefore, isn't just an attempt at modernizing processes – it is quite literally an attempt at preventing costly damages to government offices as well as citizens. And this is where identity resolution comes in.
.
In this whitepaper, you will be introduced to the concept of identity resolution, the cause of identity resolution challenges, the different identity resolution methods/solutions available and what factors to look for when choosing a vendor or software solution for your institution.
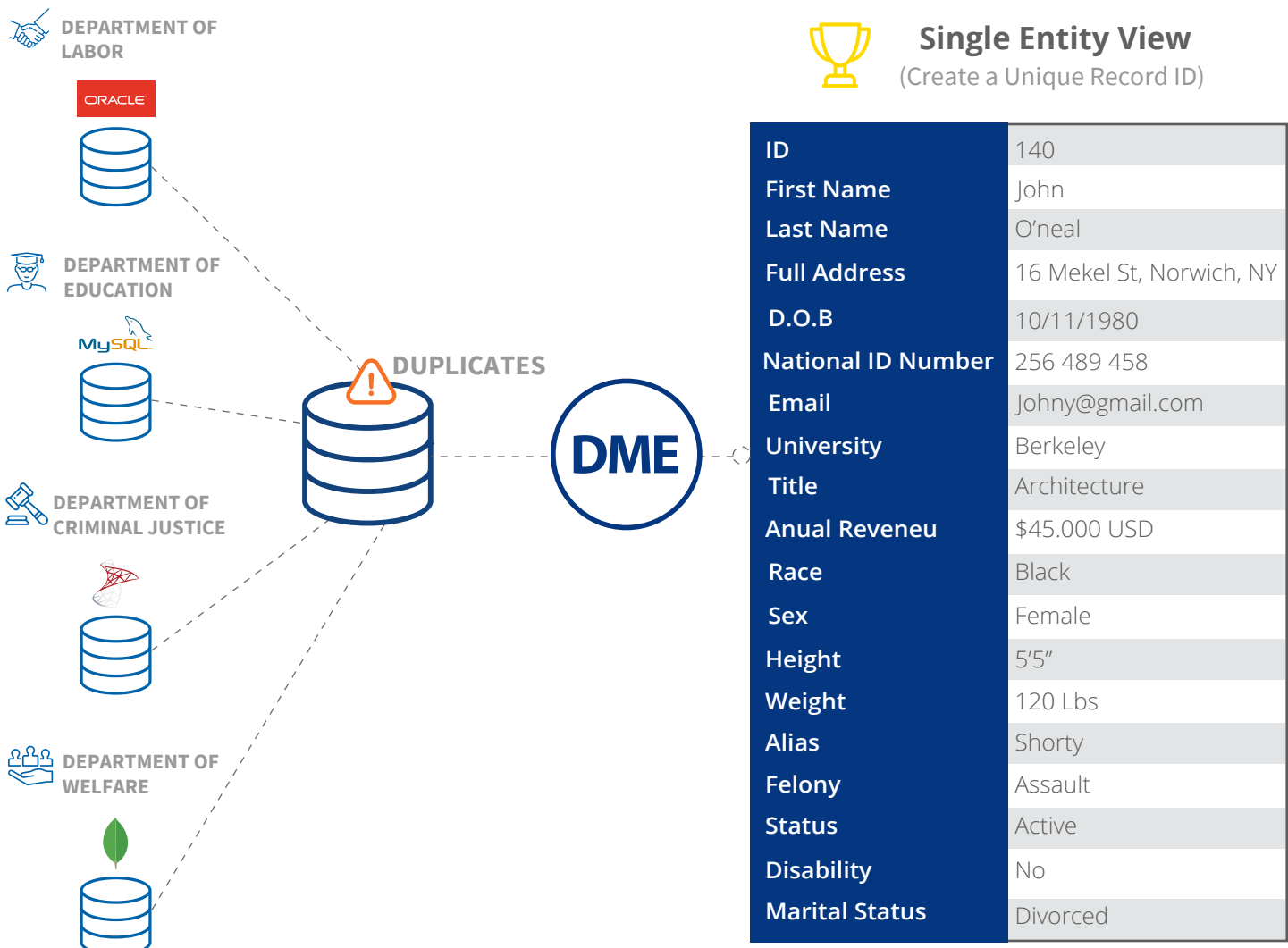
The Challenge of Unique Identities

At the most basic level, data is manually collected and stored against multiple fields such as name, age, DOB, address etc. Often, the data entry is performed by a human (either by a government officer or by the citizens themselves) which means there is always a chance of data errors. Some of the key causes of errors are:

- Typos and short forms or abbreviations entered in a hurry.
- Language barriers if names are Non-English.
- Cultural nuances in first names or surnames.
- Interpretation of abbreviations.
- Nicknames vs real names.
- Changes in contact information that was not recorded or updated properly, resulting in duplicates.

All these circumstances create a significant challenge in obtaining a unified entity view. The fundamental objective is to achieve a single source of truth or to have different instances in the same (or multiple) data sets to point to the same real-world entity.

The expectation is that a unified data asset will lead to improved processes, accurate insights and analysis, and effective communication. The search for connections between records, matching, and unification of data records is a data consolidation effort.

Data consolidation or data integration has a fundamental challenge – how do you determine if two records refer to the same entity? In rare cases, you also have to be certain that two records do NOT point to the same individual. This distinction is difficult to determine in a database riddled with duplicate or inaccurate data.

Single Entity View
(Create a Unique Record ID)

ID	140
First Name	John
Last Name	O'neal
Full Address	16 Mekel St, Norwich, NY
D.O.B	10/11/1980
National ID Number	256 489 458
Email	Johny@gmail.com
University	Berkeley
Title	Architecture
Anual Reveneu	$45.000 USD
Race	Black
Sex	Female
Height	5'5"
Weight	120 Lbs
Alias	Shorty
Felony	Assault
Status	Active
Disability	No
Marital Status	Divorced

DEPARTMENT OF LABOR

DEPARTMENT OF EDUCATION

DEPARTMENT OF CRIMINAL JUSTICE

DEPARTMENT OF WELFARE

DUPLICATES

DME

This is where the need for identity resolution comes in – a method used to find connections between records. The goal of identity resolution is to empower organizations to search, match and identify data. The methodology is applied to scan large databases across multiple jurisdictions, states, and even countries.

In the absence of unique identifiers such as SSNs, Passport numbers or other similarly sensitive information, identity resolution is a methodology that can provide an accurate view of the entity.

Identity resolution takes into its stride data profiling, data analysis and data standardization – the three critical steps in cleaning data before it actually begins comparing them. The method is performed via the use of algorithms that parse, standardize, normalize and finally compares data values.

Redundancy & Errors in Entities

For the purpose of identity resolution, let's consider an entity as a single, primary, data concept that is defined by its attributes. Common entities are, 'customers,' 'events' 'employers' etc. In a perfect world, each of these entities will have unique identifiers that can differentiate them from other instances within the dataset.

However, in the real world, there is no unique names or date of births. There are unique identifiers such as SSNs, Passport numbers, license numbers etc, but those identifiers are often confidential and are not used when data is shared.

As a simple example, consider the [name] token of an entity. Any real-world object has some kind of name that is used for reference, but the name is not unique. Cynthia Smith may also be known as Dr. Cynthia or Dr. Smith, or Dr. C.N. Smith, each name with a context in a relevant database. Cynthia's name in the Department of Public Safety may be Cynthia Naomi Smith, but in the hospital where Cynthia is working, her entry may simply be, Dr. C.N. Smith.

In any unprecedented event where the Department of Public Safety or the hospital may have to share data for analysis or research, Cynthia's record is not standardized. In this event, it may be required to determine whether Cynthia Naomi Smith, or Dr. C.N. Smith is the same person.

It is also equally possible that they are NOT the same person. Establishing these two sides of the same coin relationship is critical, especially since there may be millions of similar instances.

Although these examples sound simple, they are highly complex in nature considering that databases have billions of rows of records collected over decades. The name problem is just one of the challenges of data redundancy.

Identity resolution, therefore, tries to search and match for characteristics or attributes of any entity set to determine the match. So for example, in the case of Cynthia Smith, other attributes such as phone number, DOB, address, employer, and any other primary key are searched to determine the match.

Before Identity Resolution

1 CSUID #345
John O'neil
joneil@abccorp.com
Purchased A & D **Value = $130**

2 Web Account #0875
John
joneil@abccorp.com
Purchased B **Value = $48**

3 Call Center ID #25
Johnatan Oneal
j-oneil@hotmail.com
Returned D **Value = -$100**

4 CSUID #0345
John O'neil
j-oneil@hotmail.com
Purchased A & E **Value = $97.50**

5 No AccountID
Johnatan O'neal
j-oneil@hotmail.com
(516) 209-7869
16 Mekel St, Norwich, NY
Purchased C **Value = $68**

After Identity Resolution

Jonathan = John
NOCA Move from 156 Lane St, Fairfield CT > 16 Mekel St, Norwich, NY
Matched Cart dog food product A : Coupon purchased product A within 1 day.
joneil@abccorp.com, 3 opens, registered via email.
j-oneil@hotmail.com, 2 opens, unsubscribed

Jonathan (John) O'neal

16 Mekel St, Norwich, NY
(Moved from 156 Lane St,
Fairfield, CT

(516) 209-7869

joneil@abccorp.com (Primary)
j-oneil@hotmail.com (Secondary)
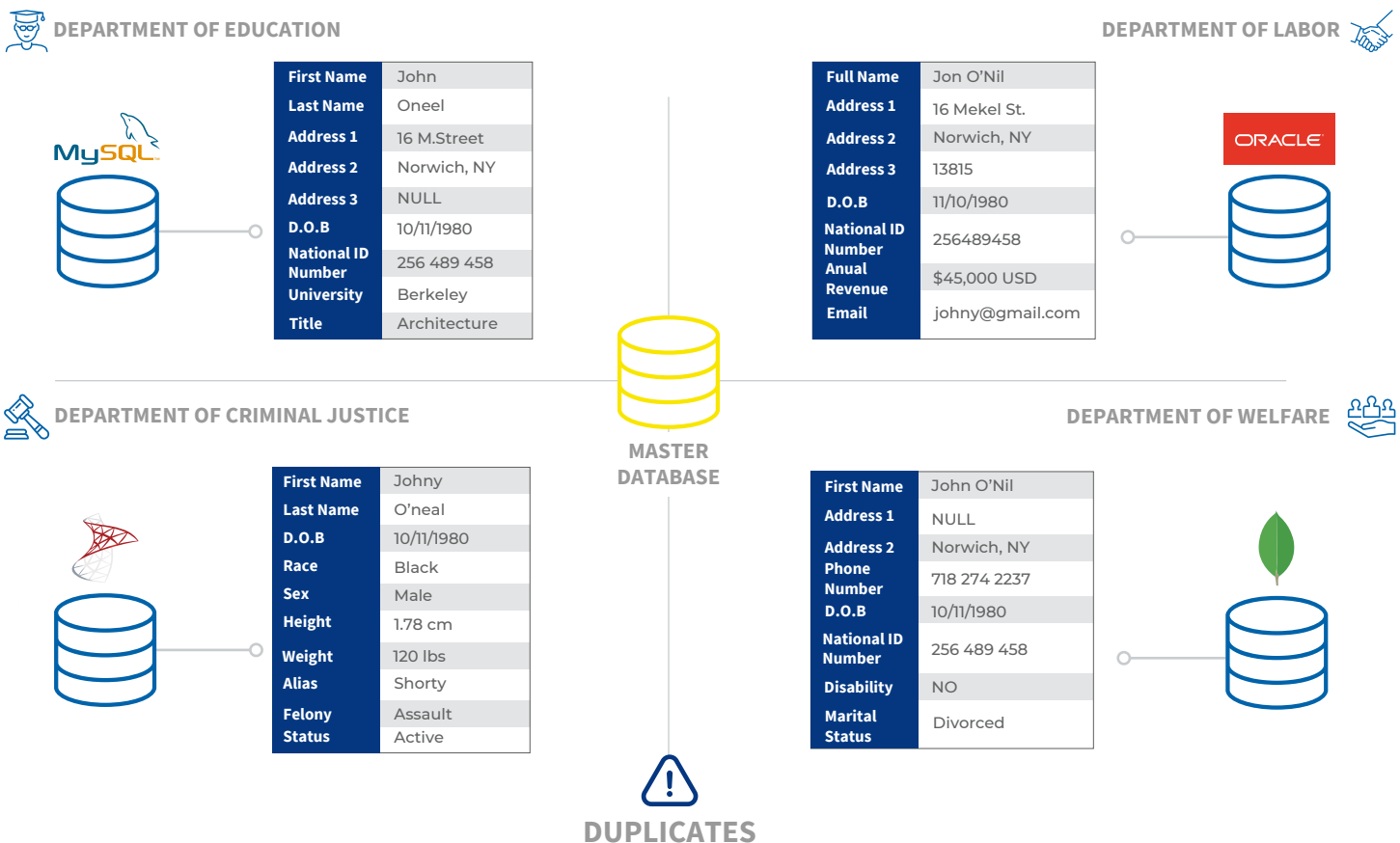
CSUID #345: Gold (Active)

$343.50
Lifetime transactions: 4
Customer since: January 15, 2017
Last transaction: October 20, 2019

Events: Call center ticket, website research, loyalty coupon.

For government organizations on a quest for a single source of truth, it's not just about identifying records – **it's about establishing the relationship of identifiers or markers to the main entity with precision.** And that's where it gets a little slippery.

Determining what values to use as identifiers is a critical task – also known formally as setting a, **'personally identifiable information' (PII) parameter.** Sometimes, such PII parameters are set if the database has such data records (registration numbers are good PIIs) or other times if the PII is confidential (such as a SSN), then a combination of all the regular attributes such as DOB, address, zip code etc are combined to get a single PII. These units on their own would not qualify as PII. Here, it is essential to note that any identity resolution solution would first parse and standardize data before the matching process. Once done, it makes use of various techniques to carry out the match.

DEPARTMENT OF EDUCATION

First Name	John
Last Name	Oneel
Address 1	16 M.Street
Address 2	Norwich, NY
Address 3	NULL
D.O.B	10/11/1980
National ID Number	256 489 458
University	Berkeley
Title	Architecture

DEPARTMENT OF LABOR

Full Name	Jon O'Nil
Address 1	16 Mekel St.
Address 2	Norwich, NY
Address 3	13815
D.O.B	11/10/1980
National ID Number	256489458
Anual Revenue	$45,000 USD
Email	johny@gmail.com

DEPARTMENT OF CRIMINAL JUSTICE

First Name	Johny
Last Name	O'neal
D.O.B	10/11/1980
Race	Black
Sex	Male
Height	1.78 cm
Weight	120 lbs
Alias	Shorty
Felony Status	Assault Active

MASTER DATABASE

DEPARTMENT OF WELFARE

First Name	John O'Nil
Address 1	NULL
Address 2	Norwich, NY
Phone Number	718 274 2237
D.O.B	10/11/1980
National ID Number	256 489 458
Disability	NO
Marital Status	Divorced

DUPLICATES

Identity Resolution Techniques

There are two ways to perform an identity resolution process – **get your in-house team of developers to build a custom solution, or use specialized data matching software**.
We'll come back later to these two choices and see which one works best for your organization. For now, let's discuss some straightforward techniques that are used in an identity resolution process.

1. Exact Match Technique: The simplest method, it uses a common identifier found on all records being sorted. This is the most basic form of identity resolution and makes use of fields such as account numbers or registration numbers that can be found across multiple databases. For records that do not require deep-level cleansing or matching, the exact match technique is the most efficient.

2. Deterministic Match Technique: Personally identifiable information (PII) such as phone numbers, emails, names, are used to identify the entity. The deterministic approach relies on the master ID that is set in the database to identify the entity. For this method to work, the master ID should remain unchanged.

3. Probabilistic Match Technique: Often discussed in line with deterministic matching, probabilistic matching is similar to deterministic matching with just a slight difference – it works on probability. If an entity has the same phone number and address across different data sets, then it is, 'probable' that it is the same entity. In this case, probabilistic matching isn't as accurate as deterministic matching, but one of the main advantages of using probabilistic over deterministic is if you want to scale your search. Say you want to match information on a third or fourth data set that may not necessarily have the deterministic identifier, then the probabilistic matching works best.

4. Similarity Match Technique: While deterministic and probabilistic methods rely on the same data values across datasets, a similarity match takes in small variations that are missed by other matching techniques. It's one of the most popular matching techniques that is accurate, scalable and can be applied to various complex datasets. For similar matching to deliver results with precision, data cleaning and data standardization is highly recommended.

All of the above techniques can be used in combination or on their own for an identity resolution objective. The effectiveness of the technique will depend on the quality of data, the number of databases and the type of datasets involved.

How Does Identity Resolution Benefit Government and Public Sector Institutes?

Unlike business databases, government databases have a direct impact on citizens' lives as well as on the operations of the state. Identity resolution can be used by any department or public institution to perform key operations such as prevention of identity fraud and ensuring public safety, research and analysis, grant raising and funding, political analysis and much more.

A quick walkthrough:

1. Public Safety and Prevention of Identity Fraud: One of the most critical advantage to identity resolution is public safety and prevention of identity fraud. Identity resolution can tap into millions of profiles and create a match of an individual based on various identifiers and data fields. For departments dealing with crime, identity fraud, immigration and other public domains, identity resolution can be used to determine attributed identity (object that links identity such as ID cards and driver license), biographical identity (social activities recorded in a public or private domain), and biometric identity (DNA/fingerprints).

2. Ensuring Compliance Against Sanction Lists: Financial institutions must comply with a number of anti-money laundering (AML) rules and regulations to ensure that they do not carry out any business or trading with businesses blacklisted in sanction lists. Identity resolution can be used to screen customer records against sanction lists such as the UN Security Council, the US Treasury Department's Office of Foreign Assets Control and the EU to prevent the possibility of money laundering.

3. Cross-Jurisdictional Research: Identity resolution is very effective for public sector institutes especially when data has to be analyzed to determine eligibility for a grant or for a research. Often, institutions in education and healthcare need cross-jurisdictional data matching for research and funding purposes. When states share data, PII fields such as SSNs, ID numbers, etc are removed. It is here when states have to rely on other primary indicators such as phone numbers and addresses to get accurate information. Identity resolution works best when working with cross-jurisdictional data.

4. Delivering Personalized Citizen Engagements: Whether it's a business or a government entity, there is an increasing need for delivering personalized engagement. Identity resolution allows for the organization to connect with citizen services in new and innovative ways. The point in concern here is citizen satisfaction with government services. Citizens want digital services initiatives that save them time, is convenient, easy to use and ideally can be performed right from their phone. In a survey, **73% of citizens** would provide biometric data such as fingerprints and retinal scans just to get more personalized services. With this level of expectation, it is only natural for government institutes to seriously consider optimizing their data in order to facilitate citizens.

5. Political Analysis and Elections: If there is one area where identity resolution is used to its maximum potential, it is in politics, especially during election times. Multiple databases are used including the National Election Pool's exit poll, the U.S. Census' Current Population Survey, the American Trends Panel and state level voter records. An analysis of all these databases gives a rich picture of a citizen's vote history, including who they are likely to support in an upcoming election. In modern times, citizen's online behavior data is also studied along with their voting data to determine the likelihood of their support for a political party. Identity resolution is so effectively used that analysts can determine which candidate a **Hispanic female is more likely to vote for.**

There really is no limit to the potential of identity resolution. All that's needed is a clean database and a powerful identity resolution tool. It's important to remember that for identity resolution to be effective, data needs to be complete, standardized and deduped.

Building a Solution In-House vs. Using an Identity Resolution Software

Restricted budgets, talent constraints and excessive overload of duties and tasks with limited staff are some of the key reasons why government and public organizations are slow on digital data growth.

Thankfully, identity resolution can now easily be performed via software solutions that are designed to handle large databases, providing data cleansing, data profiling, data standardization and a lot more.

Of course, there is always the option of either hiring data experts' in-house verses using a software, but here's a quick breakdown of the cost of hiring in-house.

Data Ladder
Get the most out of your data

If you don't build it	
Annual cost of paying for a commercial solution	**$15,000**

If you do build it	
Number of employees required	**6**
Average employee salary + benefits	**$120,000**
Weeks to build	**20**
Days per month of maintenance	**3**

$275,975+	$70,965	-$55,965	Never
Cost to build	Annual Maintenance	Saved Annually	Years to Savings

And this is just the tip of the iceberg. For government organizations, even a paid solution may be considered costly since these organizations operate on a very restricted budget. This really leaves no room for hiring in-house talent, considering the fact that most technical talents are more attracted to private-sector jobs than the government.

You can read a complete analysis of **why in house data quality solutions** fail in this extensive white paper.

That leaves us with a paid software solution. Since digital data management and data quality has been the central point of focus for over two decades now, there are plenty of automated solutions out there that can be used to accomplish identity resolution and data quality goals.

Having said that, it should be noted that not every solution provides optimal accuracy. The market standard accuracy rate of matching software solutions is 88% whereas Data Ladder's **matching accuracy is 96%** - as proven in 15 independent studies conducted by different universities.

What Should You Look for In a Commercial Software Solution?

It goes without saying that an accuracy match is the most important factor to look out when choosing a paid solution. Other than the accuracy match, there are other factors that determine whether a solution suits your needs.

1. Full-Spectrum Data Quality: Identity resolution is not just a data matching process. Before you compare data you have to ensure that it is clean. To this effect, there are three types of data cleansing functions that any market-competitive data solution must offer:

- **Data Profiling:** Most of the time, data fields are left empty or have invalid, outdated information. Data profiling is a feature that lets you sort incomplete and invalid fields of your data set and make it ready for the cleansing process.
- **Data Standardization:** New York written as NY, NYC, N.Y? This is a universal problem with most databases. To resolve this problem, you need to set automated standards to your data field. A software will let you accomplish this by automatically implementing the set standards on your data fields.
- **Data Deduplication:** This is the process where duplicates are identified and removed. Duplicates cause a significant problem for databases and are largely responsible for skewed statistics. Once data is deduped, you can get an accurate picture of your data.

ID	First Name	Last Name	Address 1	Address 2	Phone	Nat ID	State	ZIP	Birthdate
458	John	Oneel	16 M. St	Null	516465+09	256489458	Nyc	85001_	10/11/1988
2398	Jonathan	O'nil		Norw1ch	+1578698754	########	_NY		
CL_96		John O.	Mekel Street	NY	00000	25648945-8	NEW YORK	85001NY	OCT-11
Null	j0hny	O'neal	16m¶	Norwich				85001-4561	11/10/1988

Master Table

Phonetic and Typing Errors

Filled: 75%

Valid

Unstandardized Input

Leading Spaces

Trailing Spaces

Mixed Numbers and Letters

Non-Printable Characters

Incorrect Data Entry Issue

Typo

Null Values

Invalid

Inconsistent Structure

4. Data Cleansing & Standardization

Export Standardized Data | Merge Fields | Ab Proper Case Settings

Customer Master | New Prospect Records

Field Name	Type	Copy Field	Reverse Case	UPPER CASE	lower case	Proper Case	Remove Non-p... Chara...	Repla... For Non-p... Chara...	Repla... for Empty Values	Remove Leading Spaces	Remove Trailing Spaces	Chara... To Remove	Characters To Replace	Remove Spaces	Remove Letters	Remove Numbers	Replace Zeros With O Letters	Replace O Letters With...	Pattern Builder	Word Smith	Notes
RowId	#																				
Company Name	ABC																				C:\Pro...
Industry	ABC																				
Address 1	ABC																				
Address 2	ABC																				
City	ABC																				

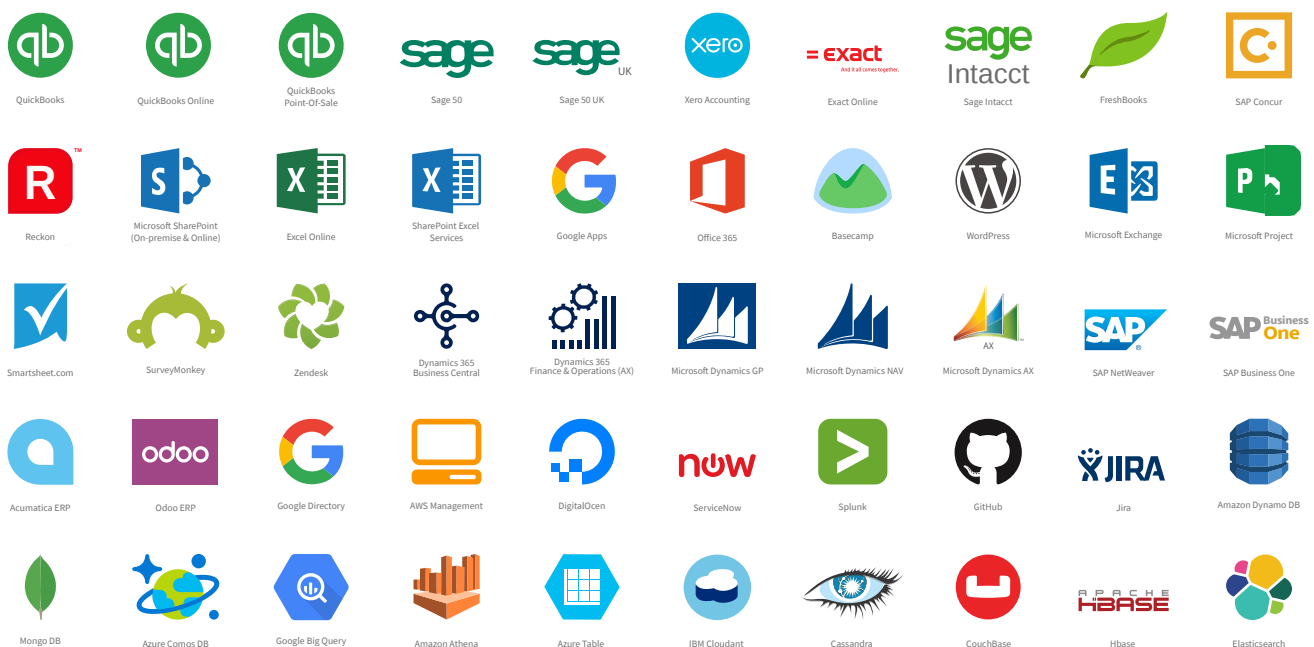*letters with numbers. and replace zeros with O's.*

Upper, Lower, Reverse or Proper casing scenarios

Remove Leading and trailing spaces, nonprintable characters & characters you don't need. For example: !"#$%&/()=,.-/*

The Pattern Builder, based on regular expressions (RegEx), is a feature that will allow users to build and search for patterns within a column and parse the match found into a new column.

Select All | Unselect All | Invert Selection | Select Field ▾

Row Id	Company Name	TC	Industry	Address 1	Address 2	City	State	ZIP	Contact Name	Prefix	First Name	Middle
2	Murphy's Fine Wine		Wines-Retail	106 Vervalen St		Closter	2613NJ	07624-2613	Vic D Jaggi		Vic	D
3	Reliance Electric		Electric Contractors	1065 W Utah Ave		Hildale	0000UT	84784-0000	Mrs Tom E Cox	Mrs	Tom	E
4	Arise Prosthetics	LLC.	Trophies Awards & Medals	Ste 101	1830 W						Mark	A
5	Brooks America		Loan brokers, nec	6 Hutton Centre Dr	# 1020						Michael	A
6	Arrowhead Jeep Tours		Tours-Operators & Promoters	841 E Paradise Ln						Mrs	Tom	B
7	Southern Utah Federal Cr Un		State credit unions	85 N 200 W							Kelly	E
8	Maez Izzy Fencing		Landscape Contractors	3000 S Yates St						Ms	Rich	D
9	GPX	INC.	Importers	900 N 23rd St							Bill	B

Count=8,487 (from 8,487)

Rules take place real time in a dynamic view

2. Integration with Systems: Be it Oracle or your own ERP – you would want to easily connect your database to the data matching software. Once you connect your database to the software you should be able to cleanse, dedupe, standardize, and enrich your data to create a 360-degree customer view, without leaving the platform.

QuickBooks | QuickBooks Online | QuickBooks Point-Of-Sale | Sage 50 | Sage 50 UK | Xero Accounting | Exact Online | Sage Intacct | FreshBooks | SAP Concur

Reckon | Microsoft SharePoint (On-premise & Online) | Excel Online | SharePoint Excel Services | Google Apps | Office 365 | Basecamp | WordPress | Microsoft Exchange | Microsoft Project

Smartsheet.com | SurveyMonkey | Zendesk | Dynamics 365 Business Central | Dynamics 365 Finance & Operations (AX) | Microsoft Dynamics GP | Microsoft Dynamics NAV | Microsoft Dynamics AX | SAP NetWeaver | SAP Business One

Acumatica ERP | Odoo ERP | Google Directory | AWS Management | DigitalOcen | ServiceNow | Splunk | GitHub | Jira | Amazon Dynamo DB

Mongo DB | Azure Comos DB | Google Big Query | Amazon Athena | Azure Table | IBM Cloudant | Cassandra | CouchBase | Hbase | Elasticsearch

3. Cross-Jurisdictional Matching Capability: As discussed before, government institutions often have to share data and conduct cross-database analysis. For this purpose, the software needs to have the capability to perform operations on multiple datasets without compromising precision and accuracy.

4. Easy to Use: Some renowned data cleansing solutions require the learning of additional language as well as trained specialists to run the software. Considering the fact that government organizations have restricted time and budget, a software that is not easy to use is counter-productive to identity resolution goals.

5. Great Customer Support: Although most data quality management tools are easy to use, you do need customer support to provide the initial deployment and training. When choosing a data solution software, it is recommended to check customer reviews and reach out to different platforms to get an idea of their customer support effectiveness.

Conclusion

As digital technology is taking over traditional methodologies, data is quickly becoming a central theme that concerns all facets of our digital society. The rapid evolution of technology accelerated by the internet, smartphones and social media has enabled global connectivity on an unprecedented scale. This connectivity has also given rise to enormous volumes of data which has made data one of the most valuable assets of the modern world.

Identity resolution, therefore, is no longer a fancy concept – it is a requirement that must be fulfilled if institutes and government departments have the objective of making the most out of their data. But there is more to identity resolution than just using matching techniques to identify data. It is an important part of a data quality management initiative – in fact, it is the first step to take if your goal is to implement a data quality management process in the coming years.

ABOUT US

Data Ladder is a data quality software company dedicated to helping business users get the most out of their data through data matching, profiling, deduplication, and enrichment tools. Whether it's matching millions of records through our fuzzy matching algorithms, or transforming complex product data through semantic technology, Data Ladder's data quality tools provide a superior level of service unmatched in the industry.

DataMatch Enterprise, was proven to find approximately 5-12% more matches than leading software companies IBM and SAS in 15 different studies.

- Unparalleled Matching Accuracy and Speed For Enterprise Level Data.

- Cleansing beating IBM and SAS.

- Proprietary Matching Algorithms with a high level of matching accuracy at blazing fast speeds on Desktop/Laptop Hardware.

- Big Data Capability

- Deduplication and Merge Purge within and across any number of files.

- Suppression of existing customers or Do Not Contact from marketing lists.

- Advanced record linking technology to create data warehouses.

- Quick Data Profile tool finds and fixes Data Quality issues within the first 5 minutes of setup to improve match quality.

Free Download